因果推論

NRIデジタル株式会社
データサイエンティスト

2020年08月08日

Share the Next Values!

ガイダンス

■ 以下の講義内容で進めていきます。

■ 質疑応答は、チャットなどでいつでも受け付けます。

- １コマ目：10時から11時半まで
  - 統計分析の復習

- ２コマ目：12時半から14時まで
  - 相関と因果の考え方
  - 効果量の測定方法
    - DIDの考え方
- ３コマ目：14時半から16時まで
  - バイアスの除去方法
  - 傾向スコアマッチング
  - IPV法
  - 効果量の測定方法
- ４コマ目：16時半から18時まで
  - 検証実験の設計
  - 検出力解析
  - 効果検証データの解析
  - 検証結果の妥当性

検証実験の設計

ABテストの設計

■ RCTを完全に満たす実験を設計することは現実的にはコストの兼ね合いから不可能。

RCTで必要となる要素

- エンドポイント　：　効果量に関する客観的指標。
- 比較対照　：　影響を受ける群と、影響を受けない群。
- ランダム化　：　集団からのランダムな抽出と、2 つの群へのランダムな割り当て。
- 盲検化　：　実験実施者と被験者は、各個人がどちらの群に属しているかを分からないようにする。

ABテストの設計

■ 以下の要素が密接に関連。

- 実験の効果量を決める。
  - どれほどの効果を見込むかを決める。

- 実験の規模を決める。
  - 効果量を測定するのに必要なサンプルサイズはどれほどか。

- 実験の対象を決める。
  - A/Bが成立するようなできるだけ似た個体が集まるグループ。
  - また、全体と被験者グループが似た傾向を持つようにする。

さらに・・・

- 実験・オペレーションコストコストの計算

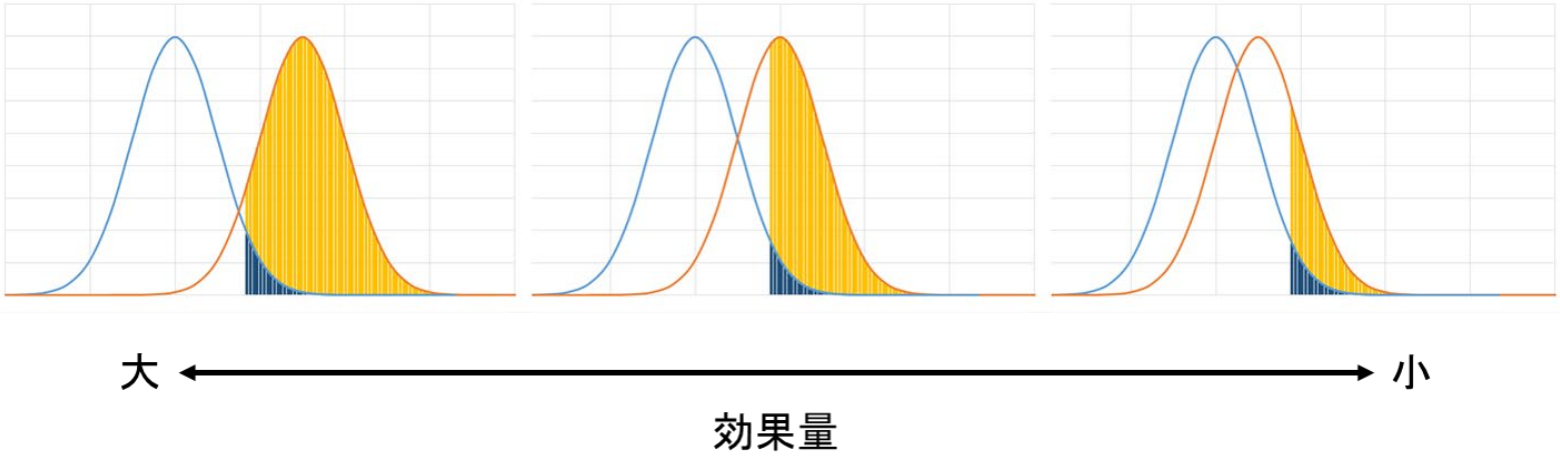
検出力解析

■ 検出力解析

- 検定では、設定した有意水準よりも p 値が小さくなったときに、分析に使用したデータに「差がある」と判断される。
- ただし、p 値はサンプルサイズが大きくなるほど反比例して小さくなる傾向があり、サンプルサイズが大きい場合有意になりやすい。
- サンプルサイズの影響を除外し、検定にかけた群に差があるかどうかを検証する方法として効果量を観察することでサンプル間の実質的な差がわかる。

- 実験計画の段階でサンプルサイズと効果量を設定しなければならない。

ある2つのグループの差を比較する実験を作りたい。　→　どれほどの効果を検出したいかを決める。

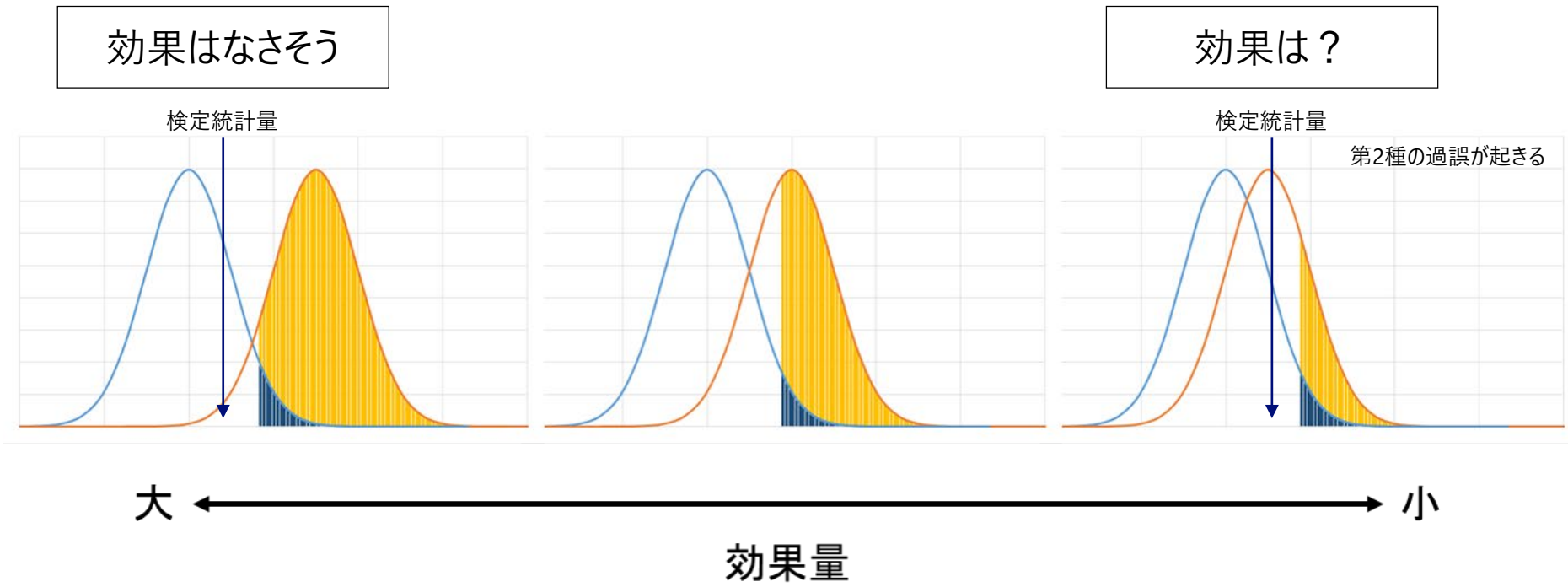


ABテストの設計

■ 検出力解析

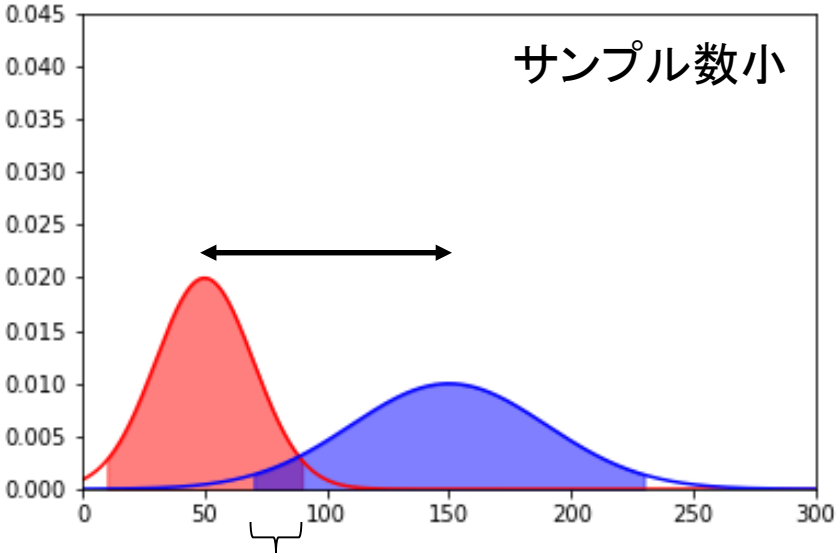
- 実験で得られる2群を想定した分布を考える。
  - 青は帰無仮説（制御群）の分布A
  - 橙は対立仮説（介入群）の分布B

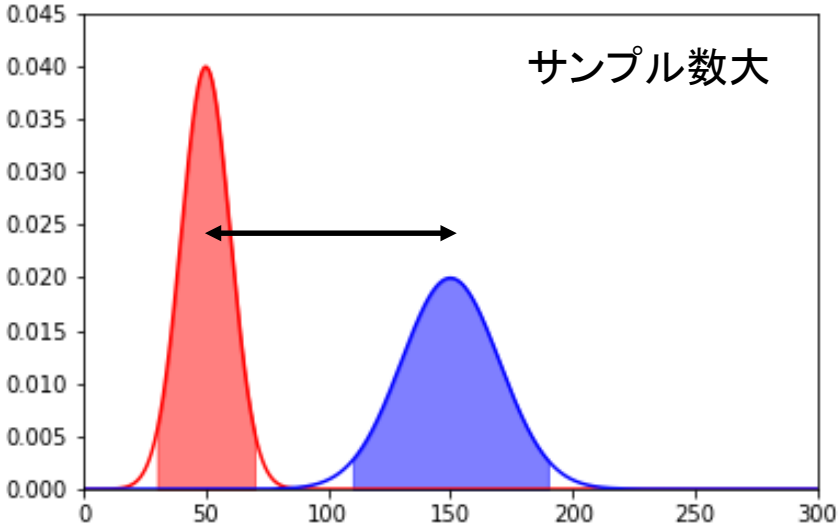
- 有意水準だけでなく、検出力（橙の面積）という。正規分布を仮定することが多いが、取得されるデータに見込まれる分布を想定することが望ましい。



サンプル数の増加で改善する様子

- 2 つのグループの間に、100の違いがあることを検定で示したい。
  - サンプル数を増やすと、それぞれのグループの推定の精度が良くなるため、分布が真の値中心に尖ってくる。
  - 2つの分布の重なりが小さくなり、2 つのグループ間の差を有意に観測することができる。
- 見込む効果量を一定とすれば、サンプル数を増加させることで改善の傾向がみられる。





分布の重なりが大きい。
→2つのグループ間の差が偶然出たものかの判別ができない

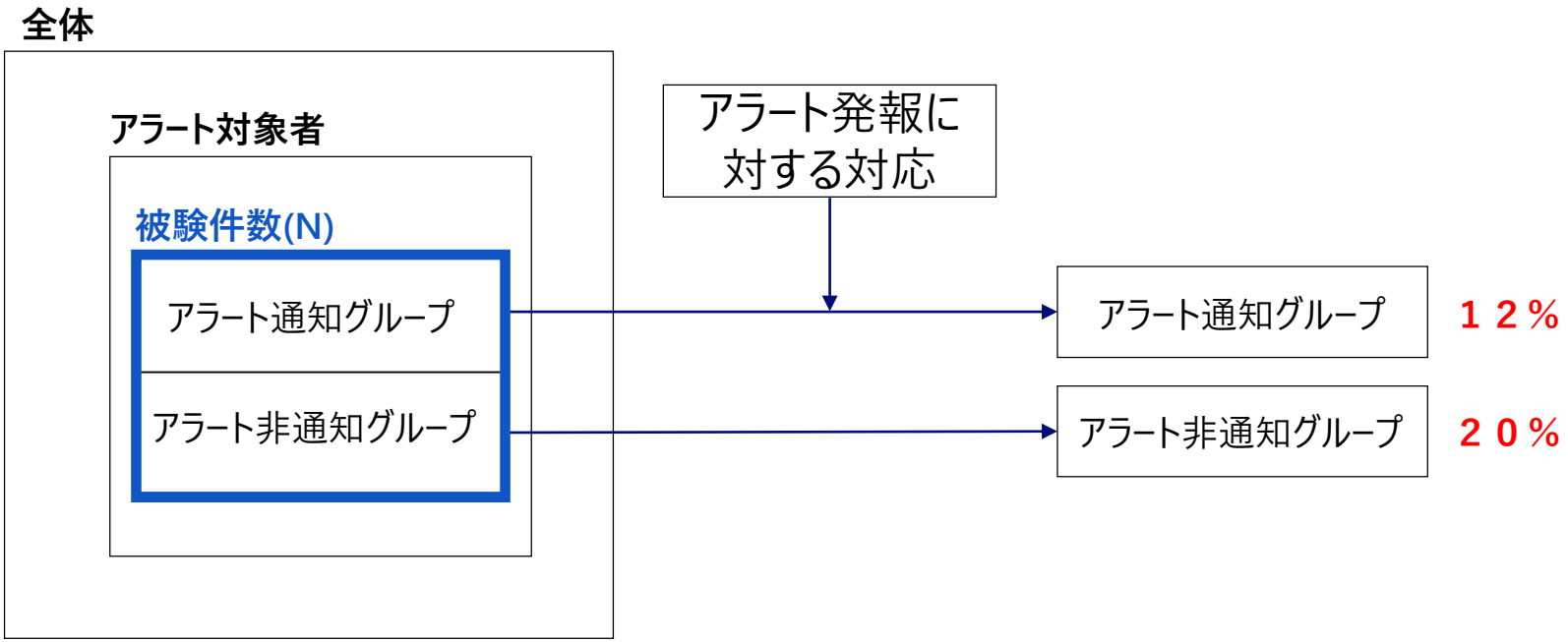
効果検証のサンプルサイズ

■ 既顧客への一定期間の疎遠を営業員にアラートとして通知する施策を全体として、比較実験を次のように行う。

- 赤字はそのグループの解約率。
- **「アラートを通知し、それに営業員が対応した」を介入**として、解約率は改善するという仮説。

開始前のアラート対象者の選別

被験グループごとの解約率の比較



効果検証のサンプルサイズ

■ 比較実験は、**サンプルサイズN、効果量Δ、有意水準α、検出力$1-\beta$**の設定を行う。

- 有意水準α =0.05、検出力$1-\beta$=0.95(80％も一般的)という一般的設定を行う。
  - 実験の結果：「差がないのに、差があるという結論」を導くのが5％。「差があるのに差がないという結論」を導くのが５％
- サンプルサイズNを定めるために、効果量を設定する（実際に、どれほど改善するのかという仮定の量）。

■ サンプルサイズの見積

- 全集団のうち、アラート対象者という性質が均質なものが対象集団である。
- ２群の比較の検定（統合比率を使ったもの）を行う。
  - 2群は対応していないが、上記の理由で、母比率は同じと考える。

■ ２群の比較の検定

- 検出力解析（２つのグループにそれぞれN必要となる仮定）
  - $N \sim \frac{2(p(1-p))(Z_{0.05}+Z_{1-0.95})^2}{\Delta^2}$

検出力95％
  - ベースとなる解約率 $p = 0.2, \Delta = 0.08, Z_{0.05} = 1.65 \rightarrow N \sim 544$

検出力80％
  - $N \sim 312$

- Nは1000あれば十分と考えられる。

データ取得後の検証

■ 2群の比較の検定（データ取得後）

- 次のようなデータを取得する（理想的なもの）。

	解約	継続
通知G	70	430
非通知G	100	400

$$p_1 = \frac{70}{500}$$

$$p_2 = \frac{100}{500}$$

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

- $Z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ を検定統計量として、α=0.05で検定する。

■ 実験における現実

- アラートを発報するも、個々人の諸事情で対応できないケース
- 十分な期間やリソースを以って対応できていないケース

  以上のようなものは**データから抜くか、非アラート組に算入**させる。アラート組の数を多く設定しておく。
- グループ間には、季節性や地域性が織り込まれているので、時期や地域性が異なるグループ設計は可能な限り避ける。

- 検定の検出力を担保するために、理論値よりも多めのデータが必要
  - 理論値はあくまでデータが理想な状態にあることが仮定されているため。

ABテストの結果の検討

■ 結果が有意でなかった場合に、どのように判断するか。

- 「実際に帰無仮説が真」なのか「単にサンプルサイズが小さくて検出力が足りないだけ」
- 複数回の実験をして確認をすることが理想。

■ 妥当性の検討

- 内的妥当性
  - 同じ実験を同じ対象に行ったら、同じ結果が再現されるだろうか。
  - 内的妥当性が失われるケース
    - 脱落変数効果
    - 関数系の不備
    - 計測誤差
    - サンプル選択
    - 双方向の因果
- 外的妥当性
  - 得られた結果を他の対象に行っても、同じ結果が得られるだろうか。
  - 母集団、地域、時期、設定が異なるときには簡単に準用してはいけない。

講義のまとめ

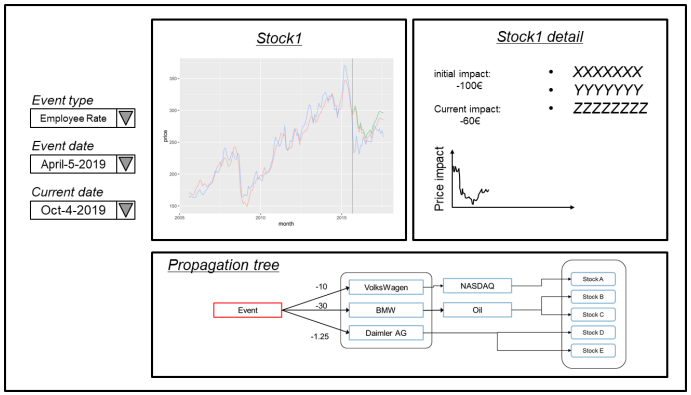
「因果推論」講義のまとめ

■ 因果推論の効果検証に関わる部分を中心にデータ解析を交えて講義を行った。

■ 因果推論は決して難しい手法を振り回す分野ではなく、データを過信せず、真摯に向き合う分野。

■ 本講義で扱った主要なトピック

- 統計分析復習
- 相関と因果
- 因果ことはじめ
- RCT
- DID
- ポテンシャルアウトカムフレームワーク
- 選択バイアス
- 傾向スコアマッチング
- 逆確率重み付け
- 中断時系列デザイン
- 不連続回帰デザイン
- 合成コントロール法



扱っていないトピック

- 因果の定義
- 因果論
- 因果構造解析
- 時系列と因果
- 機械学習と因果


Share the Next Values!